The Impact of the Pandemic on IRT Model/Data Fit

Paper presented at the annual meeting of
the National Council on Measurement in Education,

San Diego

April 9, 2022

Christie Plackner and Dong-In Kim

Data Recognition Corporation

The Impact of the Pandemic on IRT Model/Data Fit

Introduction

The application of item response theory (IRT) is almost universal in the development, implementation, and maintenance of large-scale assessments. Establishing the fit of IRT models to data is essential as the viability of calibration and equating implementations depend on it. Item level and person level fit statistics are available to diagnose and evaluate the degree of model data fit. In a typical test administration situation, measurement disturbances that influence model data fit are expected. Examples of threats to model fit include guessing, cheating (Meijer & Sijtsma, 2001), testing in an unfamiliar language, and working through the test very slowly resulting in overly ideal answers (plodding) or consequently not reaching all items (Drasgow, Levine, & Williams, 1985). Unfortunately, test administrations nationwide experienced new measurement disturbances because of the COVID-19 pandemic. Learning and associated assessments were initially suspended altogether. Return to the classroom was frequently disrupted and student learning occurred in inconsistent settings such as in-person, remote, or a combination of the two. Given the substantial disruption in education, did the response patterns of test takers change enough that model data fit is threatened and the degree of confidence in applying IRT analyses diminished?

The main purpose of this paper was to use an item fit statistic and a person fit statistic to examine the degree to which the impact of the pandemic translated to model/data misfit when IRT models are used. Model/data fit statistics for items and test takers will be evaluated for an assessment administered in 2019 and 2021. The summary item fit index Q_1 (Yen, 1993) was used as well as the person fit statistic l_z (Choi, 2010; Drasgow et. al., 1985).

Methodology

Data

Large-scale assessment data in English language arts (ELA) and mathematics grades 5 and 7 from 2019 and 2021 were used in this study. The data included item responses from test-takers with valid scores. The same mathematic forms were administered in 2019 and 2021; therefore test-takers in these years answered the same test items. The ELA forms were slightly modified. Presented item fit results consider only items common to the two years. Item types that were administered included multiple-choice (MC), multiple-select (MS), evidence-based selected response (EBSR), technology-enhanced (TE), and short

answer (SA). Table 1 summarizes the count of test takers for each administration year and the count of common items used in the analysis.

Table 1 Count of Test Takers and Items across Years

	Grade 5 ELA		Grade 7 ELA		Grade 5 Math		Grade 7 Math	
	2019	2021	2019	2021	2019	2021	2019	2021
N Students	64,531	53,793	63,683	56,013	64,631	53,711	63,830	56,053
N Common Items	36		33		46		46	

Calibration and Equating Designs

Characteristics of dichotomous and polytomous items are different, therefore, two IRT models were used in the analysis of test forms. Data were calibrated using the three-parameter (3PL model for multiple-choice (MC) items and the two-parameter partial-credit (2PPC) model for all non-MC items (Lord & Novick, 1968; Yen & Fitzpatrick, 2006). Under the 3PL model, the probability that a student with the trait or scale score θ will respond correctly to MC item *j* is

$$P_j(\theta) = c_j + (1 - c_j)/[1 + exp(-1.7a_j(\theta - b_j))].$$

In the equation, a_j is the item discrimination, b_j is the item difficulty, and c_j is the probability of a correct response by a very low-ability student. Under the 2PPC model, the probability that a student with the trait or scale score θ will respond in category k to partial-credit item j is

$$P_{jk}(\theta) = exp(z_{jk})/\sum_{i=1}^{m_j} exp(z_{ji}), \text{ where}$$
$$z_{jk} = (k-1)f_j - \sum_{i=0}^{k-1} g_{ji}, \text{ and } g_{j0} = 0 \text{ for all } j.$$

Resulting parameters estimated in the 3PL and 2PPC models are initially in two different metrics. The discrimination and location (difficulty) parameters for the MC items are in the traditional 3PL metric and are labeled *a* and *b*, respectively. In the 2PPC model, *f* (*alpha*) and *g* (*gamma*) are analogous to *a* and *b*, where *alpha* is the discrimination parameter and *gamma* over *alpha* (*g*/*f*) is the location where the adjacent trace lines cross on the ability scale. Because of the different metrics used, the 3PL parameters

a and *b* are not directly comparable to the 2PPC parameters *f* and *g*; however, they can be converted to a common metric. The two metrics are related by $b = g/f$ and $a = f / 1.7$ (Burket, 2002). As a result of this procedure, the MC and non-MC items are placed on the same scale. Note that for the 2PPC model, there are m_j–1 (where m_j is a score level *j*) independent *g*'s and one *f*, for a total of m_j independent parameters estimated for each item, while there is one *a* and one *b* per item in the 3PL model.

Calibration was performed only for spring 2019. The pre-equating design was applied to Spring 2021 due to concerns for pandemic impact on calibration and equating, and therefore spring 2019 item parameters were used for spring 2021 analyses.

Maximum likelihood estimate (MLE) was estimated as student's ability (θ).

Calibration Software

The IRT calibrations were implemented using PARDUX software (Burket, 2002). PARDUX simultaneously estimates parameters for MC and CR items using marginal maximum likelihood procedures implemented via EM algorithm.

Item Fit

Item fit was determined separately for 2019 and 2021 using the Q_1 statistic (Yen, 1993). Q_1 is a chi-square test calculated using item parameters and test takers' responses and compares observed and predicted item characteristics. Estimated item parameters and item responses were used to estimate test-taker ability (θ). These ability estimates and parameter estimates were then used to compute expected item performance. The expected item performance was compared to the observed item performance at 10 intervals across the range of test-taker achievement. The item fit statistic as computed for MC items as

$$Q_{1i} = \sum_{j=1}^{j} \frac{N_{ij}(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij}}\,,$$

where N_{ij} was the number of test takers in group *j* for item *I*, O_{ij} was the observed proportion of examinees in the cell and E_{ij} was the expected proportion of test takers. The expected proportion was computed as

$$E_{ij} = \frac{1}{N_{ij}} \sum_{a \in j}^{N_{ij}} P_i(\hat{\theta}_a),$$

where $P_i(\hat{\theta}_a$was the item characteristic function for item *I* and students *a.* The generalization of Q_1 for non-dichotomous items is

$$Gen\ Q_{ij} = \sum_{j=1}^{10} \sum_{k=1}^{mi} \frac{N_{ij}(O_{ikj} - E_{ikj})^2}{E_{ikj}}$$

$$\text{where } E_{ikj} = \frac{1}{N_{ij}} \sum_{a \in j}^{N_{ij}} P_{ik}(\hat{\theta}_a).$$

The dichotomous and generalized Q_1 statistics were transformed in *z*-scores and items were flagged if the fit statistic was > ((4*N)/1500).

The *Q1* and generalized *Q1* statistics were estimated using PARDUX (Burket, 2002; Fitzpatrick & Julian, 1996).

Person Fit

Person fit was determined separately for 2019 and 2021 test-takers. Under the IRT local independence assumption, the likelihood of observing a pattern of responses, U_j=(u_1, u_2, …, u_n), for a given examinee *j* of proficiency θ_j is expressed as following:

$$L = L(U_j|\theta_j) = \prod_{i=1}^{n} \prod_{k=1}^{m_i} P_{ik}^{u_{ik}}(\theta_j)\ ,$$

where u_{ik}=1 if the examinee chose *k* to item *i* (i.e., u_i=*k*) and u_{ik}=0 otherwise, and $P_{ik}(\theta_j)$ is the probability of endorsing category *k* of item *i* for an examinee with proficiency θ_j. The log-likelihood is then

$$ln\ L = ln\ L(U_j|\theta_j) = \sum_{i=1}^{n} \sum_{k=1}^{m_i} u_{ik}\ ln\ P_{ik}\ (\theta_j)\ .$$

Since $ln\ L$ is dependent on θ, Drasgow et al. (1985) proposed a standardized log-likelihood statistic (l_z) as follows (Choi, 2010):

$$l_z = \frac{ln\ L - E(ln\ L)}{\sqrt{Var(ln\ L)}},$$

where $E(ln\ L)$ denotes the expected value of $ln\ L$

$$E(ln\ L) = \sum_{i=1}^{n} \sum_{k=1}^{m_i} P_{ik}\ ln\ P_{ik}\ (\theta_j)\ ,$$

and $Var(ln\ L)$ denotes the variance of $ln\ L$

$$Var(ln\,L) = \sum_{i=1}^{n}\left[\sum_{k=1}^{m_i}\sum_{h=1}^{m_i} P_{ik}(\theta)P_{ih}(\theta)\,ln\,P_{ik}\,(\theta)\,ln\frac{P_{ik}}{P_{ih}}\right].$$

Because the value of the l_z statistic decreases as the extent of person misfit, a low likelihood of observing response pattern U_j=(u_1, u_2, …, u_n), increases, large negative values of the statistic indicate misfitting item response vectors. In practice, the unknown parameter θ is replaced by $\hat{\theta}$ in the computation of l_z. Drasgow et al. (1985) found that the empirical distribution of $l_z$was close to the standard normal distributions for long tests with more than 80 items, and van Krimpen-Stoop and Meijer (1999) found that the use of $\hat{\theta}$ in place of θ may result in a decreased variance of $l_z$for short tests.

l_z statistic and cutoff values were estimated by adding the 3PL model to PERSON$_Z$ (Choi, 2010). The 2PPC item parameters for the non-MC items were transformed to the item parameters based on the generalized partial credit model (GPCM) for PERSON$_Z$. MLE ranges for simulation were restricted between -4 and 4 as can be seen in Figure 1. Cutoff values were obtained by simulation with number of simulees, 100,000, and the 3PL/GPCM model. The mean and standard deviation for simulation were those of the empirical values of MLEs. As an example, Figure 1 illustrates where the critical values lay within the Grade 7 mathematics empirical and simulated data distributions. The x-axis is the estimated thetas and the y-axis is the l_z statistic. The critical value for α = 0.05 is represented by the red dashed lines and the α = 0.01 critical value is represented by the blue dashed line.

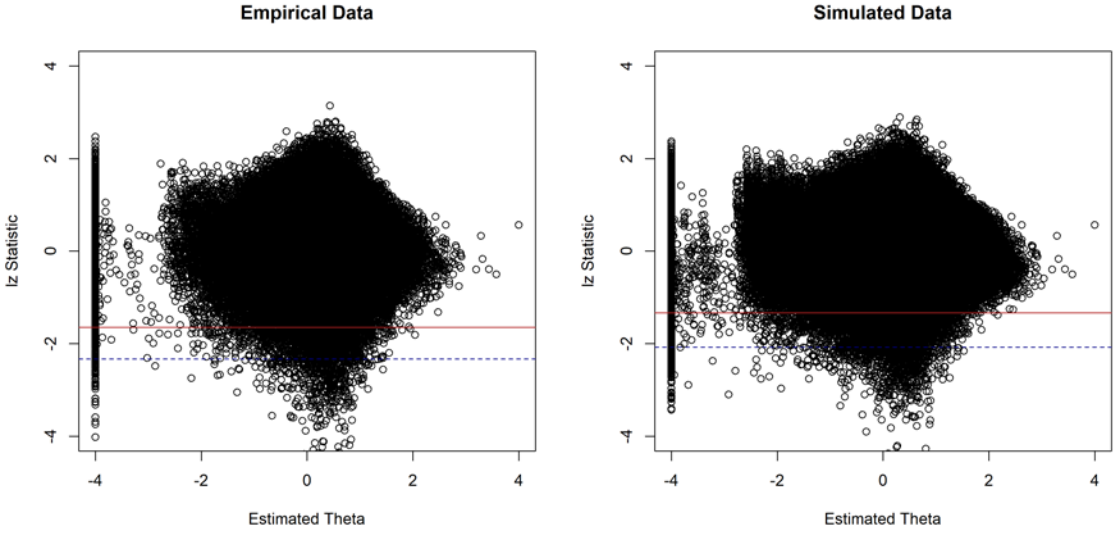

Figure 1 Visualization of Critical Values using Empirical and Simulated Data for α = 0.05 (red dashed line) and α =0.01 (blue dotted line)

Figure 2 illustrates how person fit looks when plotted. The example includes four test-takers; the title on each panel provides MLE ability and the *p*-values associated for the empirical $l_z(\theta)$. The *p*-value was included instead of the l_z for ease of interpretation. For example, if the *p*-value was less than 0.05, the test taker would be flagged when the critical value for α = 0.05 was used. Similarly, if the *p*-value was less than 0.01 the test taker would be flagged when the critical value for α = 0.01 was used. In Figure 2, the lower two panels would be flagged for either critical value situation. The x-axis presents the items in increasing order of average score; therefore, items around 0 are the difficult items and items near 40 are the easy items. The y-axis is the student's average score for each item. The top two panels illustrate item score patterns for test takers not flagged for person fit; the plots allow one to see that their item scores trends towards 1 as the items become easier, even though there were some exceptions. Alternatively, the lower panels illustrate patterns for test-takers flagged for misfit, the item response/score pattens seems more random with a score of 0 on some easy items and a full score (1) on some difficult items.

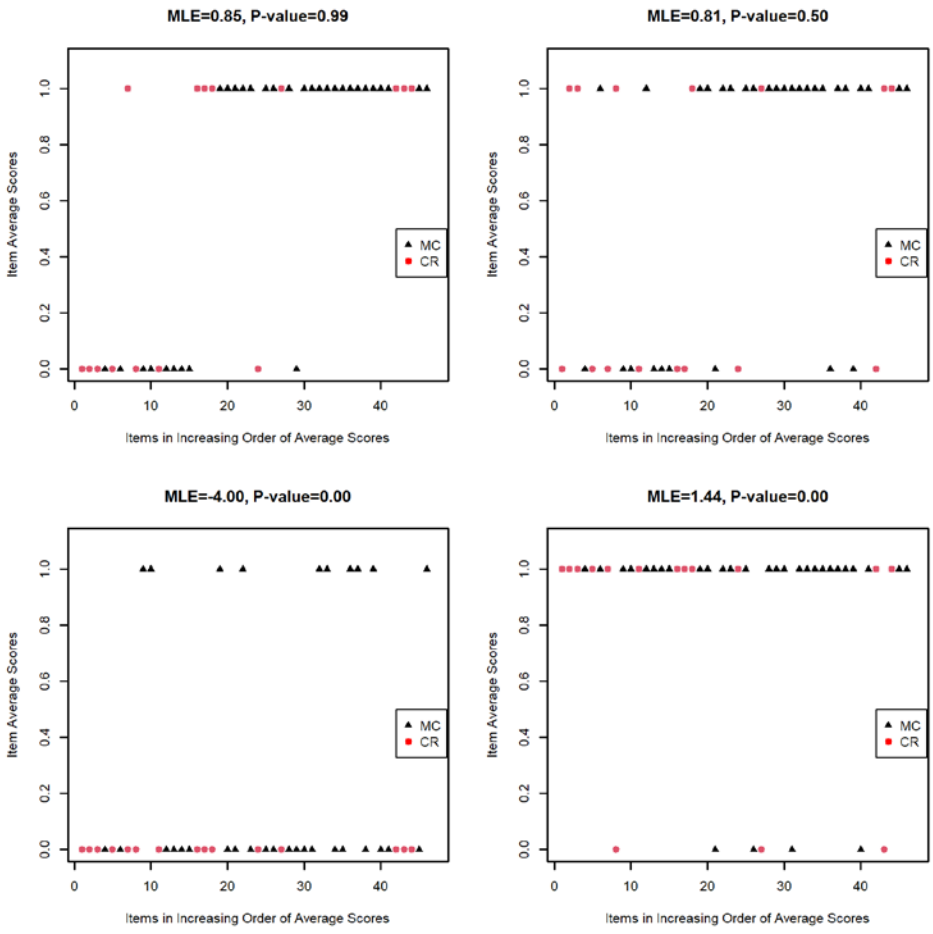

Figure 2 Sample of Test-Taker Person Fit Plots for Math Grade 7

Once test-takers were examined for person fit, those that were flagged were summed to the district level. Groups were examined at the district level to determine if the proportion of flagged test-takers increased in 2021 from 2019.

Results

Item Fit

Table 2 and Table 3 summarize the count of common items and flagged items for ELA and Table 4 and Table 5 do so for mathematics. The count of flagged items was examined at an overall level and also by items' type, and content domain.

Although the ELA tests were slightly modified in 2021 from the 2019 administration, there were 36 common items administered in Grade 5 and 33 items in Grade 7. In Grade 5 ELA, of the 36 common items, three items were flagged for item misfit in 2019 and two in 2021 and no item was flagged in both 2019 and 2021. Two of the flagged 2019 items were multiple-choice (MC) items and one was a technology-enhanced (TE) item; one of the items was aligned to Domain 1 and two to Domain 3. Both flagged 2021 items were MC; one was a Domain 1 item and the other a Domain 3 item. Grade 7 ELA saw more items flagged for item fit than Grade 5, five in 2019, seven in 2021, and zero items flagged both years. There were 18 common MC items administered and in both 2019 and 2021, three of them were flagged for misfit (although not the same three items). Two and four TE items were flagged in 2019 and 2021, respectively.

The intact 2019 mathematics forms were administered in 2021 so all items were common across the years. In Grade 5 mathematics, of the 46 common items, two were flagged in 2019 and four were flagged in 2021 and two of the items were flagged both years. In 2019 Grade 5 there was one flagged item each for the item types short answer (SA) and TE. One item aligned to Domain 2 and the other to Domain 5. Grade 5 saw items of all item types flagged in 2021: one MC, one SA, and two TE. Two of the items were in Domain 2 and two in Domain 5. Of the items flagged for item fit in both years, one was an SA item and the other a TE and they aligned one each to the content domains of 1 and 5. Grade 7 mathematics also had 46 common items and two were flagged in 2019, five were flagged in 2021, and one was flagged in both years. The 2019 flagged items were of item types SA and TE and of content domains 1 and 5. There were two MC items, one SA item, and two TE items flagged in 2021. Two items were aligned to Domain 1, one Domain 3, and two Domain 4. The item that was flagged in both years was a TE item in Domain 1.

Table 2 Summary of Common Items in 2019 and 2021 and Items Flagged for Item Fit: Grade 5 ELA

		N Common Items	N Flagged in 2019	N Flagged in 2021	N Flagged Both Years
	Total	36	3	2	0
Item Type	MC	24	2	2	0
	MS	1	0	0	0
	EBSR	1	0	0	0
	TE	10	1	0	0
Item Domain	1	6	0	0	0
	2	18	1	1	0
	3	12	2	1	0

Table 3 Summary of Common Items in 2019 and 2021 and Items Flagged for Item Fit: Grade 7 ELA

		N Common Items	N Flagged in 2019	N Flagged in 2021	N Flagged Both Years
	Total	33	5	7	0
Item Type	MC	18	3	3	0
	MS	2	0	0	0
	EBSR	4	0	0	0
	TE	9	2	4	0
Item Domain	1	5	1	1	0
	2	16	3	4	0
	3	12	1	2	0

Table 4 Summary of Common Items in 2019 and 2021 and Items Flagged for Item Fit: Grade 5 Mathematics

		N Common Items	N Flagged in 2019	N Flagged in 2021	N Flagged Both Years
	Total	46	2	4	2
Item Type	MC	27	0	1	0
	SA	12	1	1	1
	TE	2	1	2	1
Item Domain	1	9	0	0	0
	2	10	1	2	1
	3	9	0	0	0
	4	9	0	0	0
	5	9	1	2	1

Table 5 Summary of Common Items in 2019 and 2021 and Items Flagged for Item Fit: Grade 7 Mathematics

		N Common Items	N Flagged in 2019	N Flagged in 2021	N Flagged Both Years
	Total	46	2	5	1
Item Type	MC	31	0	2	0
	SA	7	1	1	0
	TE	8	1	2	1
Item Domain	1	10	1	2	1
	2	10	0	0	0
	3	7	0	1	0
	4	8	0	2	0
	5	11	1	0	0

Person Fit

Table 6 summarizes the distribution of the person fit statistic l_z, reporting the mean, standard deviation, minimum value, and maximum value. The critical values used for flagging items are also reported.

Table 6 Summary of Person Fit l_z Statistic Distribution

	Grade 5 ELA		Grade 7 ELA		Grade 5 Math		Grade 7 Math	
	2019	2021	2019	2021	2019	2021	2019	2021
Mean	0.07	0.02	0.05	0.01	0.05	0.02	0.07	0.09
Std Dev	0.75	0.76	0.79	0.81	0.82	0.81	0.27	0.88
Min	-4.25	-4.09	-3.44	-4.02	-4.26	-4.45	-5.80	-4.60
Max	3.48	3.02	3.08	3.07	2.96	2.91	3.14	2.94
Critical Value α = 0.05	-1.2110	-1.2052	-1.2784	-1.2914	-1.2977	-1.3019	-1.3326	-1.3636
Critical Value α = 0.01	-1.9056	-1.9055	-1.9861	-1.9913	-2.0361	-2.0334	-2.0748	-2.1082

Test takers were evaluated for person fit using two critical values; one determined using α = 0.05 and one at α = 0.01 based on simulation. For each year, the percentage of flagged test takers was determined for each level at the district level. Districts were flagged if 10% or more of the test takers were flagged for person misfit. In addition to having at least 10% of the test takers flagged, the district test-taker count had to be at least 10. Table 7 summarizes the total count of test takers, the percentage flagged at each level and the difference of percentage flagged between 2021 and 2019. Additionally, the table summarizes the count of districts that were represented in the year's data, the count of districts that had at least 10 test takers, and the percentage of districts flagged for having at least 10% of the test takers flagged. If the percentage of flagged districts was larger in 2019 than 2021, the font is red.

Except for Grade 7 math, the percentage of test takers flagged increased from 2019 to 2021. The change in percentage of test takers flagged at the α = 0.05 level ranged from -0.84% in Grade 7 math to 0.82% in grade ELA. The percentage of change at the α = 0.01 was even smaller, ranging from -0.28% at Grade 7 math to 0.31% at Grade 5 ELA.

The percentage of districts flagged for person misfit at the α = 0.05 level increased in 2021 for all grades and subjects, except for Grade 7 math. The largest change in flagged districts was seen at Grade 5 ELA with 4.26%. Essentially no change in percentage of districts flagged was seen at the α = 0.01 level as the range of change was -0.19% in G7 ELA to 0%.

Table 7 Summary of Test Taker and District Counts and Percentage of Flagged Test Takers

			% of Test Takers Flagged		Districts		% Districts Flagged	
		N Test Takers	α 0.05	α 0.01	N Total	N to be Flagged	α 0.05	α 0.01
G5 ELA	2019	64,531	4.81	0.94	561	530	7.36	0.38
	2021	53,793	5.63	1.25	578	525	11.62	0.38
	2021-2019	-10,738	0.82	0.31	17	-5	4.26	0.00
G7 ELA	2019	63,683	5.29	1.08	559	523	8.22	0.19
	2021	56,013	5.97	1.30	575	527	11.20	0.00
	2021-2019	-7,670	0.68	0.22	16	4	2.97	-0.19
G5 Math	2019	64,631	5.77	1.25	561	530	13.02	0.57
	2021	53,711	5.95	1.28	578	526	14.64	0.57
	2021-2019	-10,920	0.18	0.03	17	-4	1.62	0.00
G7 Math	2019	63,830	6.63	1.51	559	523	17.97	0.76
	2021	56,053	5.79	1.23	575	527	13.85	0.76
	2021-2019	-7,777	-0.84	-0.28	16	4	-4.12	-0.01

The percentage of test takers flagged relative to various demographic groups was also summarized, see Table 8 to Table 11. If the percentage flagged is greater in 2019 than 2021, then the font is red. Although patterns can be seen as to which groupings had a larger percentage of flagged students – for example, across all years and subjects a larger percentage of males were flagged for person misfit than females – a clear pattern of changes in percentages was not evident. The most consistent change was within the SES demographic group, those who are not SES disadvantaged had a slightly larger percentage of increased test takers flagged than those that were SES disadvantaged. In ELA, although small, the percentage of students flagged among the groups was consistently larger in 2021 than 2019. This is not

true in Grade 7 mathematics where every group, except for the ethnicity category of Asian had more test takers with person misfit in 2019 than 2021.

Table 8 Percentage of Test Takers Flagged for Person Misfit: Grade 5 ELA

		2019		2021		2021-2019	
		α	α	α	α	α	α
Demographic Group*		0.05	0.01	0.05	0.01	0.05	0.01
State	State	4.81	0.94	5.63	1.25	0.81	0.30
Ethnicity	Hispanic	5.75	1.04	6.51	1.66	0.76	0.63
	Black	6.65	1.12	7.50	1.56	0.85	0.44
	White	4.32	0.90	5.21	1.12	0.89	0.22
	Asian	4.25	0.81	5.25	0.98	1.00	0.17
	Other	5.20	1.02	6.42	1.66	1.22	0.64
Gender	Female	3.62	0.66	4.26	0.90	0.64	0.25
	Male	5.97	1.22	6.94	1.58	0.97	0.36
EL	No	4.65	0.91	5.48	1.23	0.84	0.33
	Yes	6.88	1.41	7.70	1.46	0.82	0.05
TTS	No	4.08	0.76	4.87	1.02	0.78	0.27
	Yes	8.25	1.81	9.57	2.41	1.33	0.59
Disability	No	4.23	0.81	5.02	1.07	0.80	0.26
	Yes	9.00	1.91	9.87	2.49	0.87	0.58
SES	No	3.85	0.80	4.84	1.06	0.99	0.26
	Yes	6.02	1.13	6.89	1.55	0.87	0.42
PL	1	7.75	1.41	8.61	1.88	0.86	0.47
	2	6.95	1.61	7.87	2.01	0.92	0.41
	3	1.30	0.09	1.59	0.09	0.29	0.00
	4	0.00	0.00	0.00	0.00	0.00	0.00

*EL = English Learner; TTS = Text-to-Speech; SES = Socio-economic Disadvantaged; PL = Proficiency Level

Table 9 Percentage of Test Takers Flagged for Person Misfit: Grade 7 ELA

		2019		2021		2021-2019	
		α	α	α	α	α	α
Demographic Group*		0.05	0.01	0.05	0.01	0.05	0.01
State	State	5.29	1.08	5.97	1.30	0.68	0.21
Ethnicity	Hispanic	6.25	1.24	7.22	1.49	0.97	0.25
	Black	6.51	1.31	7.30	2.26	0.78	0.95
	White	4.90	1.02	5.52	1.14	0.63	0.13
	Asian	4.89	0.95	6.56	1.23	1.67	0.28
	Other	5.88	1.21	6.71	1.60	0.83	0.39
Gender	Female	4.78	0.94	5.91	1.29	1.14	0.35
	Male	5.79	1.22	6.03	1.30	0.24	0.08
EL	No	5.20	1.06	5.80	1.29	0.61	0.23
	Yes	6.91	1.45	9.03	1.48	2.12	0.03
TTS	No	4.97	1.01	5.64	1.24	0.67	0.23
	Yes	7.41	1.55	8.36	1.73	0.95	0.18
Disability	No	5.02	1.06	5.71	1.26	0.70	0.19
	Yes	7.33	1.24	8.03	1.63	0.70	0.39
SES	No	4.53	0.94	5.40	1.14	0.87	0.20
	Yes	6.35	1.29	6.97	1.57	0.62	0.29
PL	1	6.86	1.21	7.31	1.45	0.45	0.24
	2	8.22	2.14	9.37	2.43	1.14	0.30
	3	2.95	0.31	3.19	0.40	0.24	0.09
	4	0.23	0.00	0.22	0.00	-0.01	0.00

*EL = English Learner; TTS = Text-to-Speech; SES = Socio-economic Disadvantaged; PL = Proficiency Level

Table 10 Percentage of Test Takers Flagged for Person Misfit: Grade 5 Mathematics

		2019		2021		2021-2019	
		α	α	α	α	α	α
Demographic Group*		0.05	0.01	0.05	0.01	0.05	0.01
State	State	5.77	1.25	5.95	1.28	0.19	0.03
Ethnicity	Hispanic	5.85	1.31	5.95	1.20	0.10	-0.12
	Black	5.38	1.20	6.25	1.12	0.88	-0.08
	White	5.91	1.25	5.94	1.30	0.03	0.04
	Asian	4.29	1.19	4.90	1.18	0.61	-0.01
	Other	5.66	1.14	6.43	1.53	0.77	0.40
Gender	Female	5.54	1.12	5.68	1.17	0.14	0.06
	Male	5.99	1.37	6.22	1.38	0.23	0.01
EL	No	5.80	1.26	5.93	1.29	0.13	0.03
	Yes	5.33	1.12	6.25	1.19	0.92	0.06
TTS	No	5.56	1.19	5.83	1.23	0.26	0.05
	Yes	6.71	1.51	6.62	1.51	-0.09	0.00
Disability	No	5.43	1.14	5.72	1.19	0.29	0.05
	Yes	8.16	2.01	7.64	1.93	-0.52	-0.08
SES	No	5.52	1.14	5.76	1.24	0.24	0.10
	Yes	6.08	1.39	6.27	1.34	0.19	-0.04
PL	1	5.54	1.08	5.28	0.99	-0.26	-0.09
	2	6.26	1.54	6.24	1.37	-0.02	-0.17
	3	6.83	1.50	7.60	1.82	0.77	0.32
	4	1.75	0.07	1.57	0.04	-0.18	-0.03

*EL = English Learner; TTS = Text-to-Speech; SES = Socio-economic Disadvantaged; PL = Proficiency Level

Table 11 Percentage of Test Takers Flagged for Person Misfit: Grade 7 Mathematics

		2019		2021		2021-2019	
		α	α	α	α	α	α
Demographic Group*		0.05	0.01	0.05	0.01	0.05	0.01
State	State	6.63	1.51	5.79	1.23	-0.84	-0.28
Ethnicity	Hispanic	5.51	1.21	5.00	0.96	-0.51	-0.24
	Black	6.96	1.49	5.30	0.79	-1.66	-0.70
	White	6.81	1.57	6.01	1.32	-0.80	-0.25
	Asian	6.46	1.45	6.50	1.51	0.03	0.06
	Other	6.64	1.70	4.91	1.08	-1.72	-0.61
Gender	Female	5.80	1.24	4.96	0.95	-0.84	-0.28
	Male	7.41	1.77	6.57	1.50	-0.84	-0.28
EL	No	6.63	1.50	5.82	1.24	-0.81	-0.26
	Yes	6.54	1.63	5.21	1.05	-1.33	-0.57
TTS	No	6.60	1.54	5.82	1.26	-0.77	-0.28
	Yes	6.82	1.31	5.51	1.04	-1.30	-0.28
Disability	No	6.54	1.51	5.75	1.22	-0.79	-0.29
	Yes	7.27	1.52	6.07	1.29	-1.20	-0.23
SES	No	6.65	1.55	5.90	1.31	-0.75	-0.24
	Yes	6.59	1.46	5.58	1.09	-1.00	-0.37
PL	1	6.00	1.08	5.07	0.90	-0.93	-0.18
	2	7.40	1.88	6.16	1.48	-1.24	-0.40
	3	7.22	1.80	6.68	1.49	-0.54	-0.31
	4	1.86	0.13	1.59	0.00	-0.27	-0.13

*EL = English Learner; TTS = Text-to-Speech; SES = Socio-economic Disadvantaged; PL = Proficiency Level

Summary and Discussion

Every test administration has some expected measurement disturbances, such as, but not limited to, guessing, cheating, test anxiety, or plodding. However, conditions created by the pandemic to the learning and testing environment of large-scale assessment test takers could have produced conditions that had magnified or introduced new measurement disturbances. Item fit and person fit indices were used to examine if model data fit was a larger concern in 2021 than in 2019 since IRT was applied to calibration, equating, and scoring.

Item fit was determined separately for 2019 and 2021 using the Q_1 statistic (Yen, 1993). Items that were common to both administrations were examined to see if more items were flagged in 2021 than in 2021. While one more item was flagged in 2019 for Grade 5 ELA, more items were flagged in 2021 for Grade 7 ELA and grades 5 and 7 mathematics. However, the count of additional flags was quite

small with the largest increase being three items in Grade 7 mathematics. Items were grouped by item type and by content domain to determine if a type or domain was more likely to include items presenting misfit in 2021 than 2019, but a pattern was not seen.

Person fit was examined using the l_z statistic (Choi, 2010; Drasgow et. al, 1985) to determine if more test takers were flagged for misfit in 2021 than 2019. Similar to item fit, the increase in flagged test takers were low and for Grade 7 mathematics the percentage of flagged students was higher in 2019. For the other grades and subjects the largest increase in flagged test takers was in Grade 5 ELA with 0.82% at the $\alpha = 0.05$ level. Districts were flagged if 10% of more of the students in the district were flagged for misfit. At the $\alpha = 0.05$ the percentage of districts flagged increased by 1.62% in Grade 5 math to 4.26% in Grade 5 ELA but decreased by 4.12% in Grade 7 math. Various demographic groups were reviewed to see if the percentage of students flagged with the groups shifted from 2019 to 2021, other than a small increase overall, except for Grade 7 math, the shifts were all small with the largest percentage increase in Grade 7 ELA where the percentage of flagged students increased by 2% for English learners.

Overall, there does not seem to be a greater threat to the feasibility of using IRT models in 2021 than previous years. However, this study was limited to the use of one item fit and one person fit index. It is also important to note that while the forms administered in 2021 were essentially the same (with small changes to ELA) as those administered in 2019, the population was not. There were approximately 11,000 fewer test takers in Grade 5 and 8,000 fewer in Grade 7 in 2021 than in 2019. The greatest effect of the COVID-19 pandemic could have been experienced by these missing test takers.

References

Choi, S (2010). PERSON$_z$: Person Misfit Detection using the l_z Statistic with Monte Carlo Simulations. *Applied Psychological Measurement*, 34(6), 457-458.

Burket, G. R. (2002). PARDUX [Computer program]. Unpublished.

Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-68.

Fitzpatrick, A. R., & Julian, M. W. (1996). *Two studies comparing the parameter estimates produced by PARDUX and PARSCALE*. Unpublished manuscript.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading,

MA: Addison-Wesley.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159–176.

van Krimpen-Stoop, E.M.L.A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, *23*, 327-345.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.

Yen, W. M., & Fitzpatrick, R. R. (2006). Item response theory. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger Publishers.